# Proper Noun Recognition in Cross-Language Record Linkage by Exploiting Transliterated Words

Yuting Song
Taisuke Kimura
Biligsaikhan Batjargal
Akira Maeda

Ritsumeikan University, Japan

RITSUMEIKAN

# Background

- The same data entity can exist in different languages across data sources.

Database in Japanese

| | |
|---|---|
| 番号 | 92202743 |
| 作品名/資料名 | **冨嶽三十六景 神奈川沖浪裏** |
| 作家/制作者 | 葛飾北斎/画 |
| 制作年 | 天保2年～4年 |
| 所蔵館 | 江戸東京博物館 |

| | |
|---|---|
| 番号 | 92202746 |
| 作品名/資料名 | **冨嶽三十六景 深川万年橋下** |
| 作家/制作者 | 葛飾北斎/画 |
| 制作年 | 天保2年～4年 |
| 所蔵館 | 江戸東京博物館 |

| | |
|---|---|
| 番号 | 08200001 |
| 作品名/資料名 | **雪月花 隅田** |
| 作家/制作者 | 葛飾北斎/画 |
| 制作年 | [天保3年] |
| 所蔵館 | 江戸東京博物館 |

Database in English

**Under the Wave off Kanagawa (Kanagawa oki nami ura), also known as The Great Wave, from the series Thirty-six Views of Mount Fuji (Fugaku sanjūrokkei)**
Katsushika Hokusai (Japanese, Tokyo (Edo) 1760–1849 Tokyo (Edo))

Date: ca. 1830–32

Medium: Polychrome woodblock print; ink and color on paper

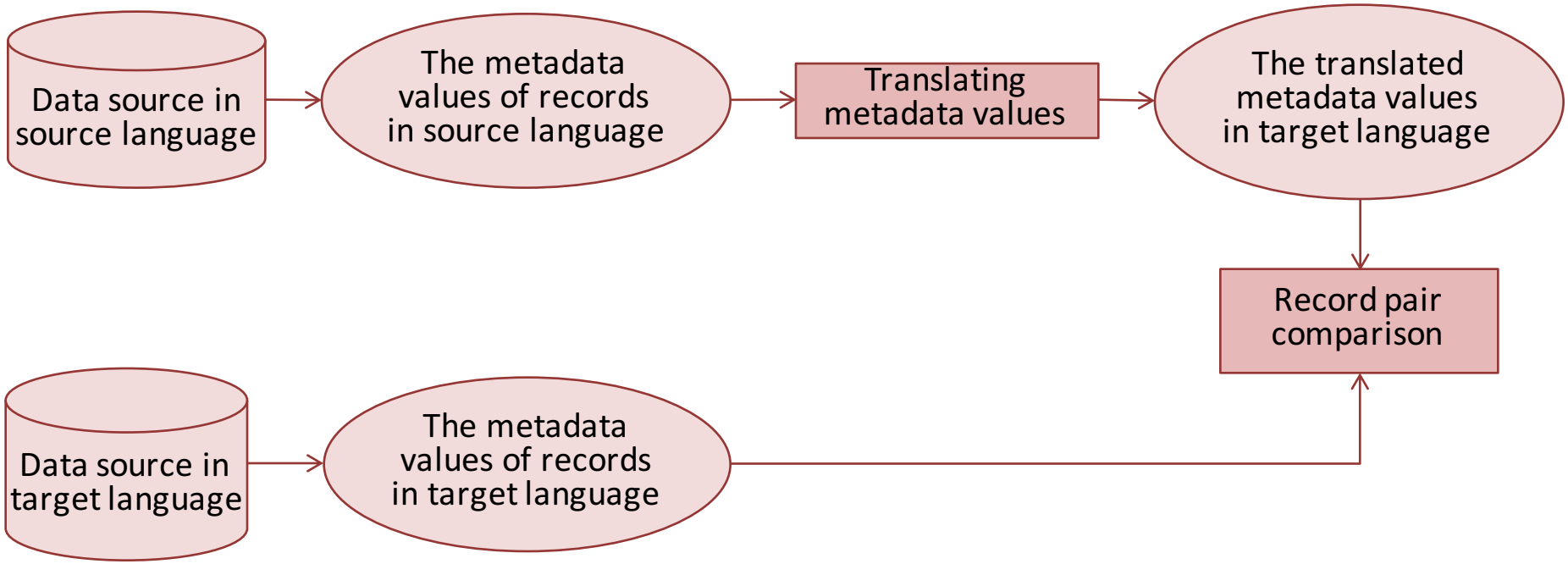
**Snow on the Sumida River (Sumida), from the series, Snow, Moon, and Flowers (Setsugekka)**
Katsushika Hokusai (Japanese, Tokyo (Edo) 1760–1849 Tokyo (Edo))

Date: ca. 1833

Medium: Polychrome woodblock print; ink and color on paper

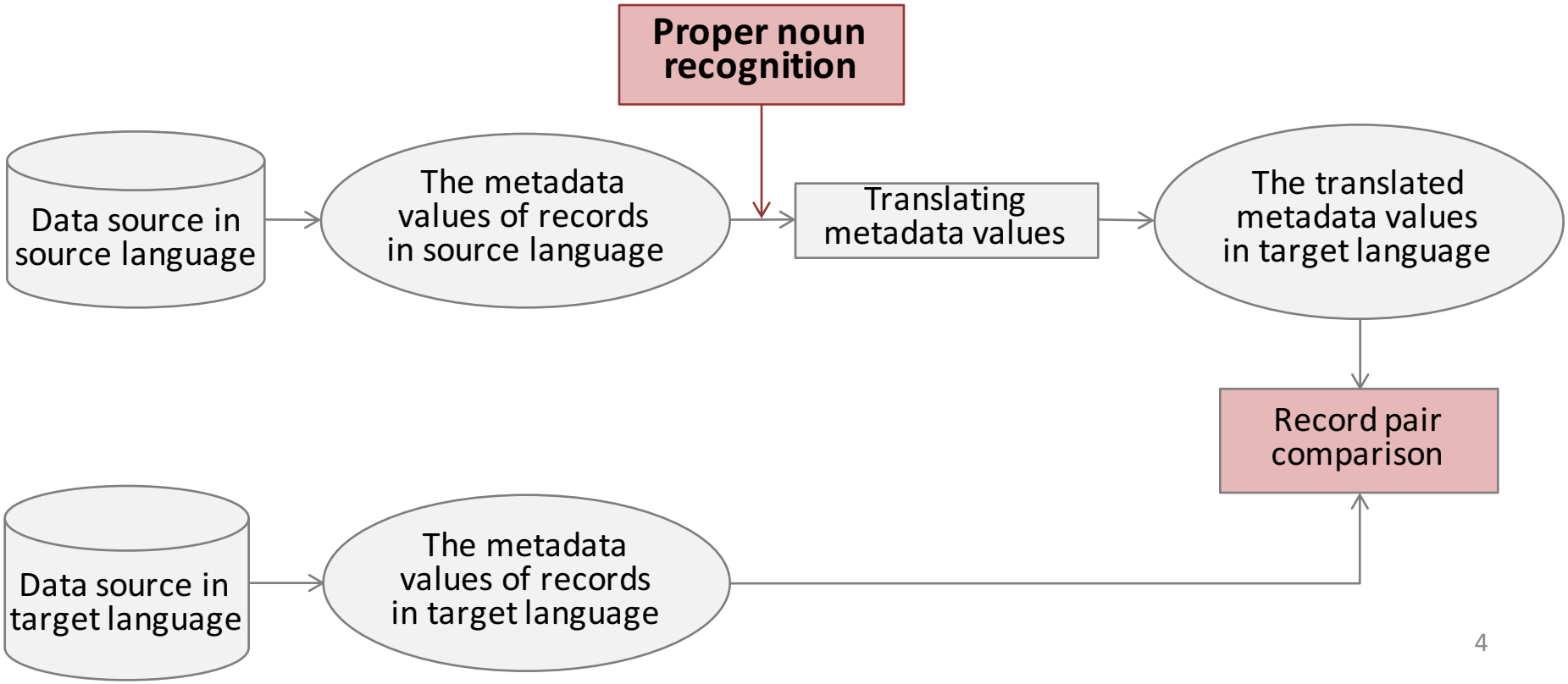
# Motivation

- **Cross-language record linkage**
  - **Record linkage** is to find record pairs that refer to the same entity across multiple data sources
  - **across multiple data sources in different languages**
  - It provides opportunities for people to access multilingual information



# Problem statement

- **Proper nouns** in metadata values are more easily to be matched
- **Correct recognition and translation** of proper nouns in metadata values could have a positive effect on cross-language record linkage



# Problem statement

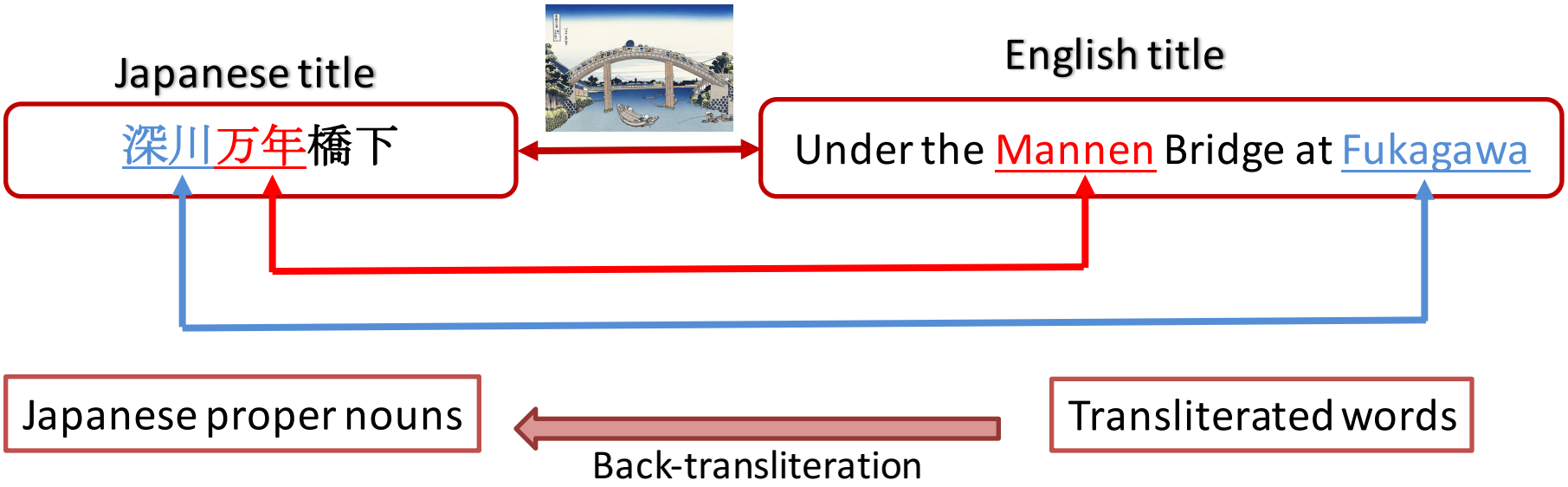
- **Proper nouns** in metadata values are more easily to be matched
- **Correct recognition and translation** of proper nouns in metadata values could have a positive effect on cross-language record linkage

**How to recognize proper nouns in metadata values?**

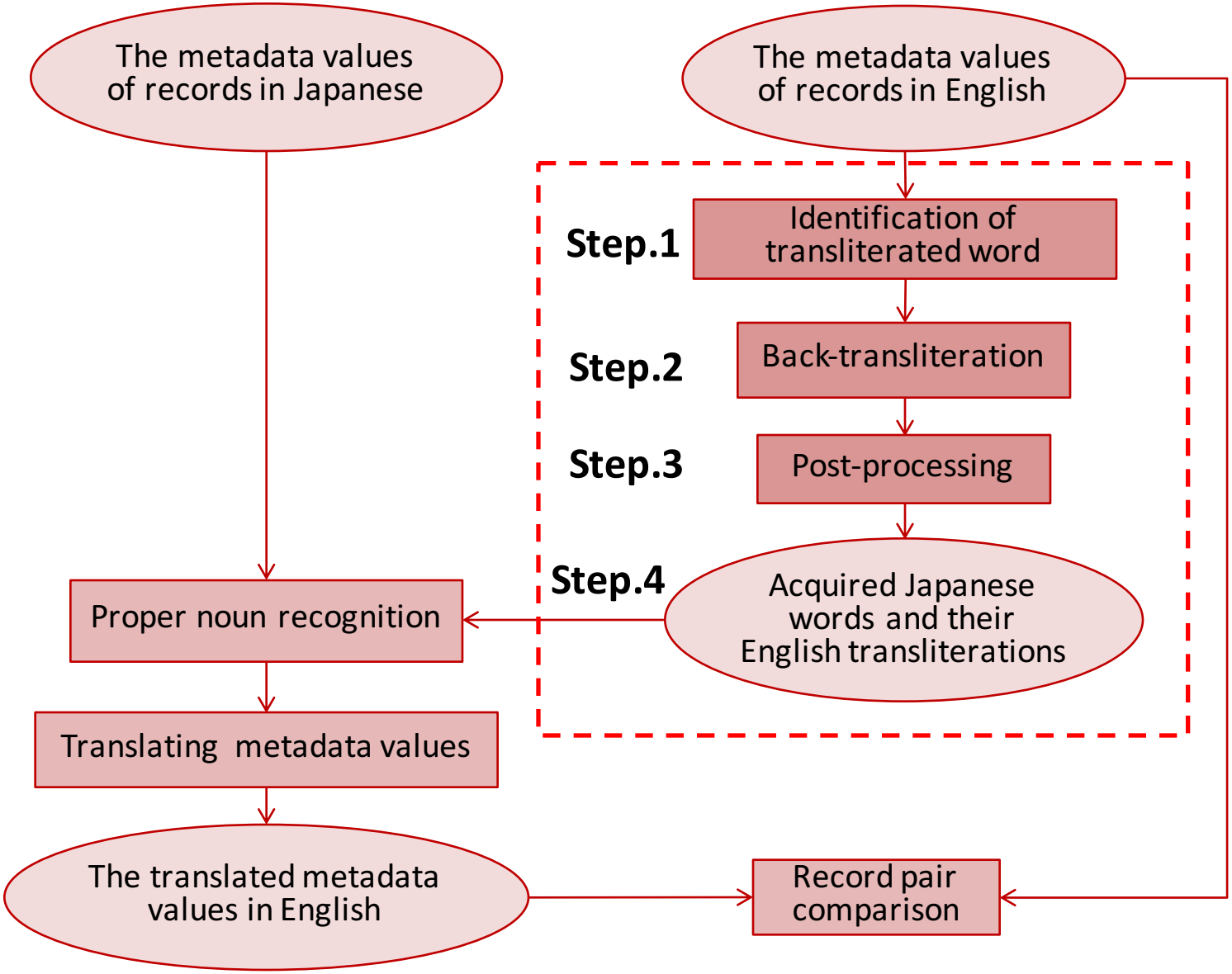
- **Named Entity Recognition system**
  - It does not perform well on metadata values
    - The metadata values are usually short texts that can't provide sufficient grammar and syntax information

# Proposed method

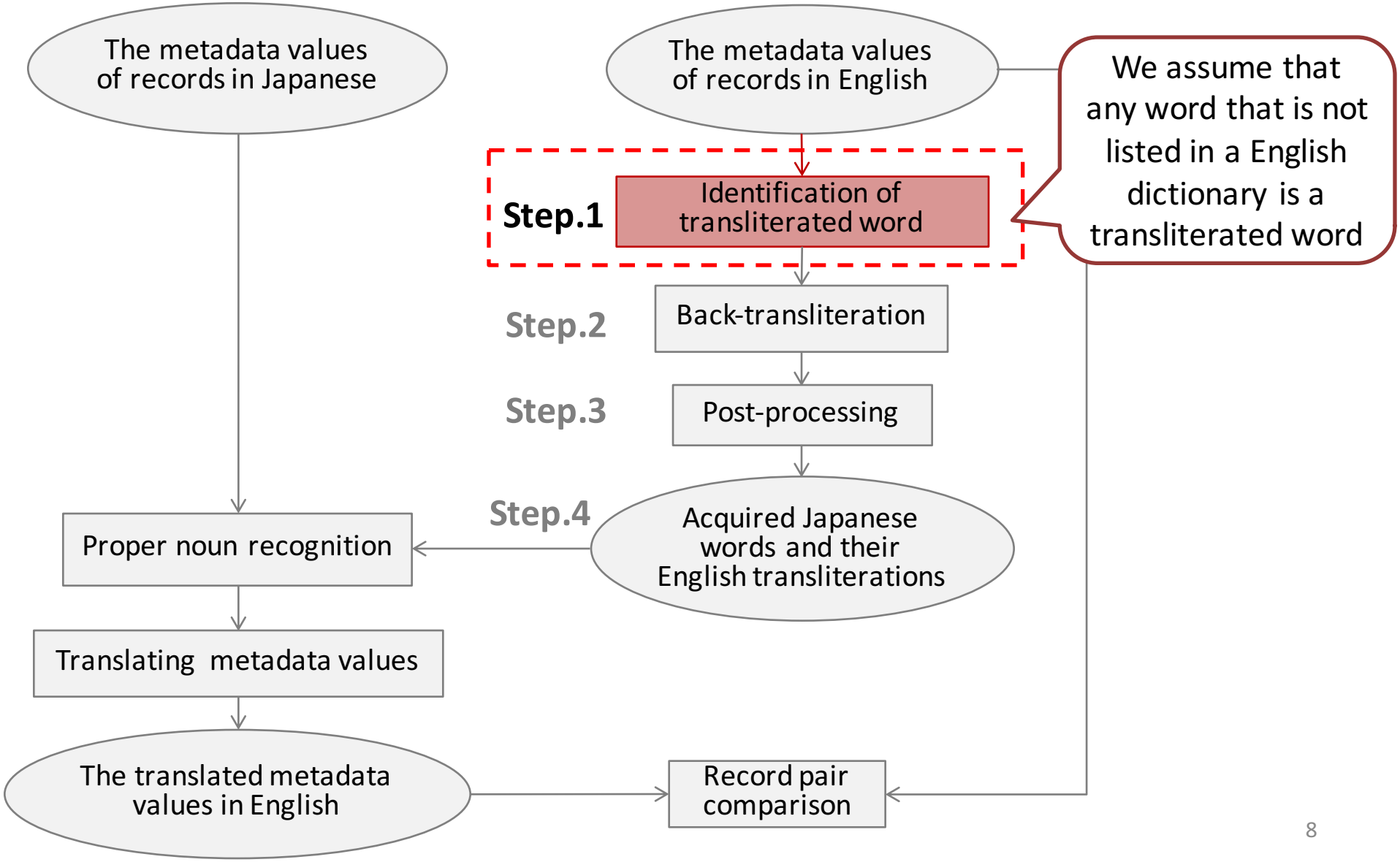
- Our proposed method is inspired by the following observation: **the English translation of a Japanese proper noun is usually a transliterated word**
- English transliterated words are easy to be identified
- **Back-transliteration**: converting English transliterated words to their corresponding Japanese words



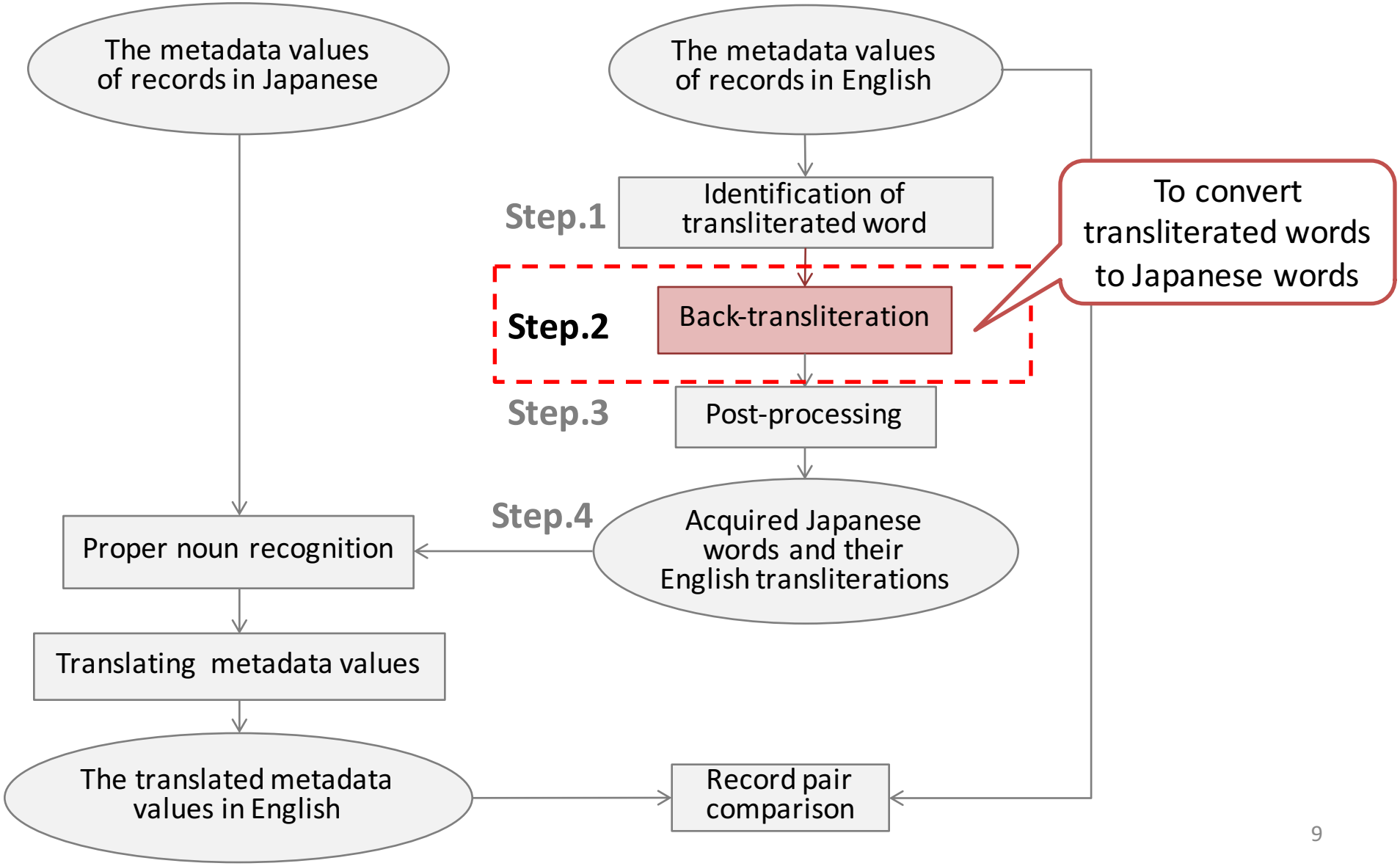
# Overall process



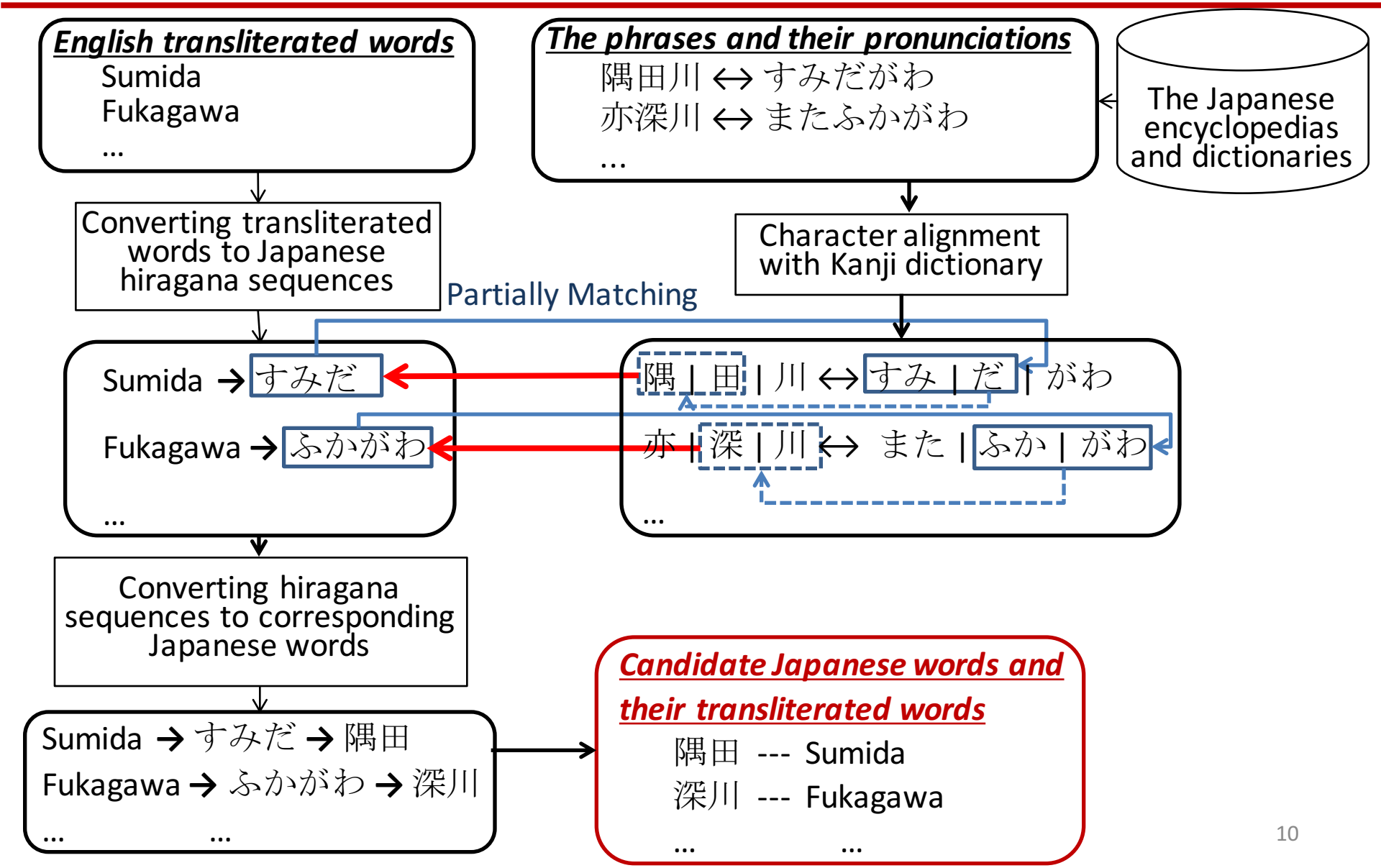
# Identification of transliterated word



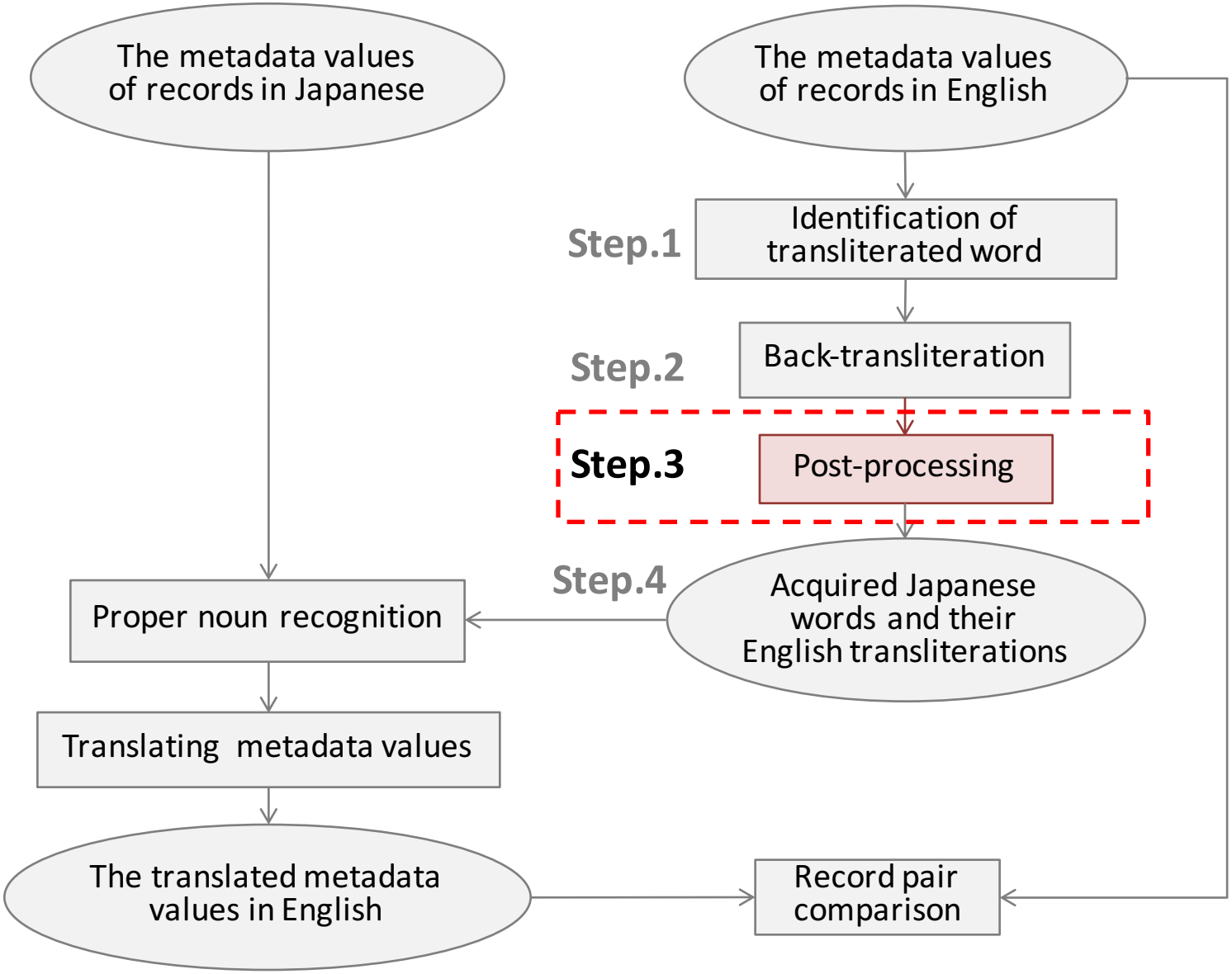
# Back-transliteration



# Back-transliteration

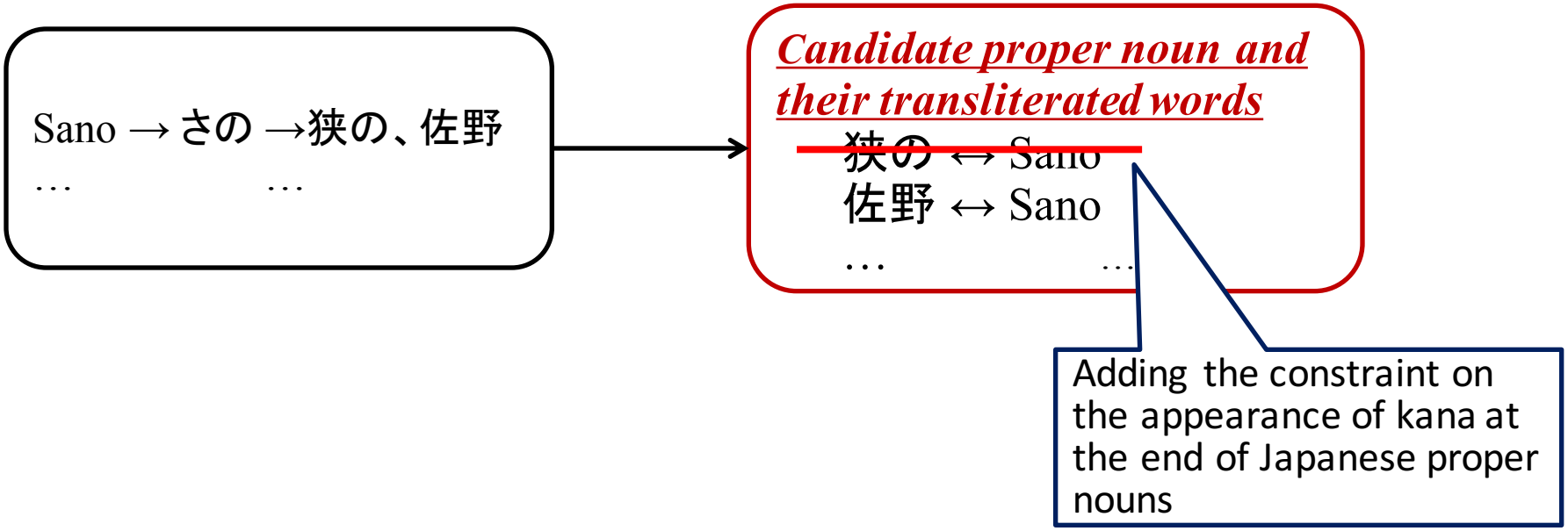


# Post-processing



# Post-processing

- One English transliterated words could **be back-transliterated to one or more Japanese words**
  - Some of them are not proper nouns
- Post-processing is to remove the words that are not proper nouns



# Experiments

- Linking the same Ukiyo-e records between databases in Japanese and English
- Experimental data
  - Ukiyo-e prints



    - Japanese traditional woodblock printing
    - These prints have been digitized and exhibited in many digital libraries and museums with textual metadata values in various languages
  - Dataset
    - The titles of Ukiyo-e prints are used to identify the same records

| Language | Ukiyo-e database | Number of Ukiyo-e prints |
|---|---|---|
| Japanese | Edo-Tokyo Museum | 242 |
| English | Metropolitan Museum of Art | 3456 |

    - Each Japanese title has at least one corresponding English title
    - Among the 242 Japanese titles, 209 titles contain at least one proper noun (*Target titles*)

# Cross-language record linkage

- **Recognition of proper nouns** in metadata values
  - **Our proposed method**
  - **Baseline**: MeCab – a Japanese part-of-speech and morphological analyzer
    - An example of using MeCab to recognize proper nouns

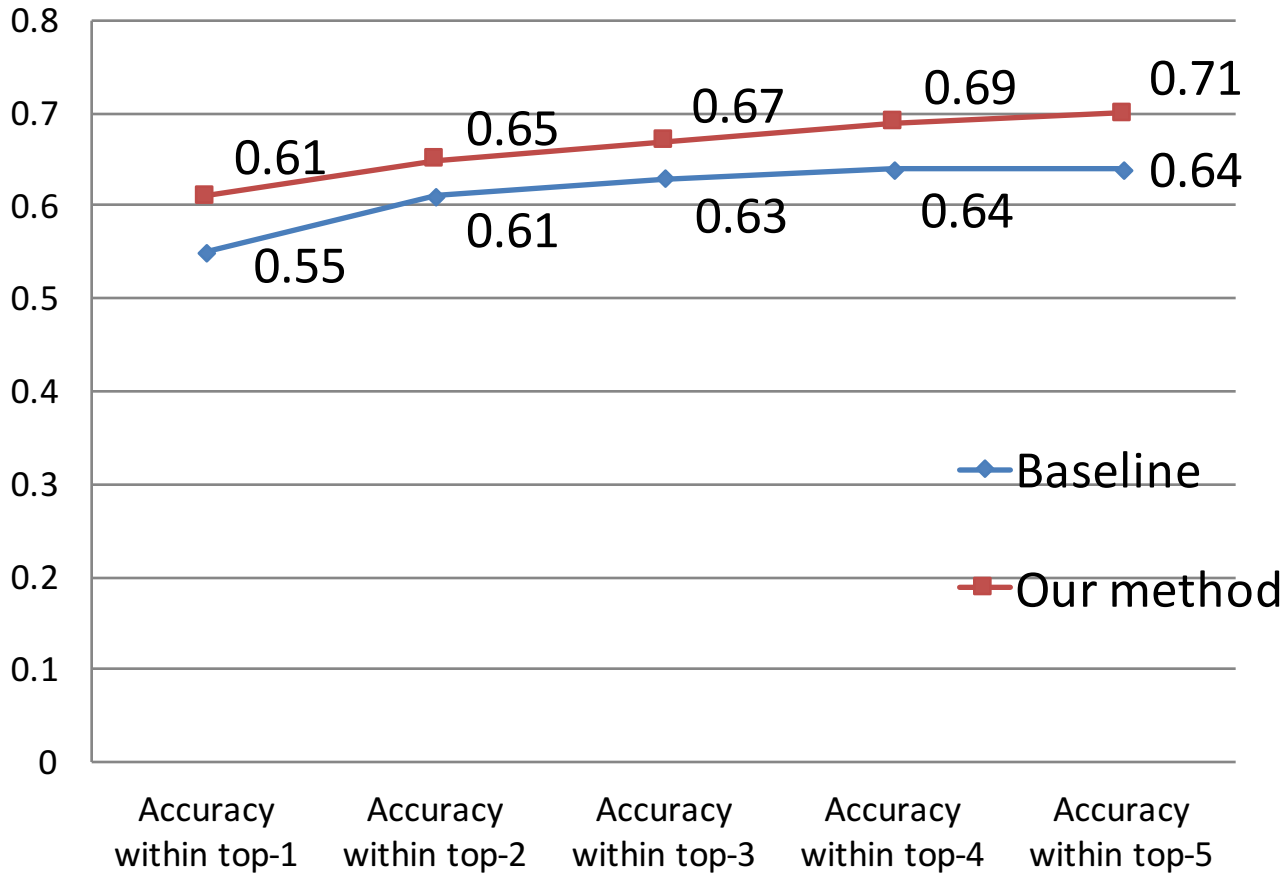
Input: 深川万年橋下

Output:

| Word | Part-of-speech | Sub-class of part of speech | Pronunciation |
|---|---|---|---|
| 書字形（＝表層形） | 品詞 | 中分類 | 発音形出現形 |
| 深川 | 名詞-固有名詞-地名-一般 | 固有名詞 | フカガワ |
| 万 | 名詞-数詞 | 数詞 | マン |
| 年 | 名詞-普通名詞-助数詞可能 | 普通名詞 | ネン |
| 橋下 | 名詞-普通名詞-一般 | 普通名詞 | キョーカ |

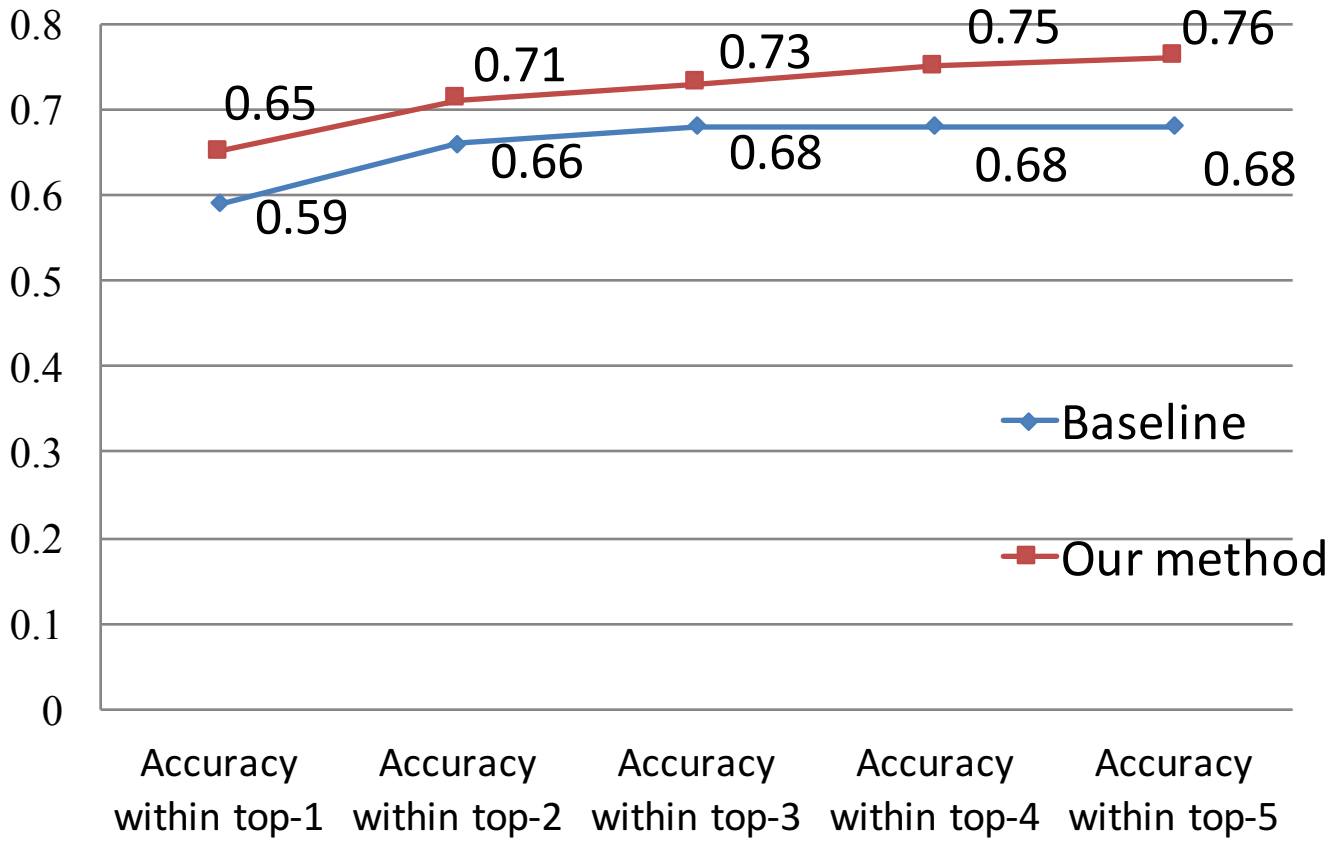
# Experimental results (1)

- Linking the same Ukiyo-e records between databases in Japanese and English



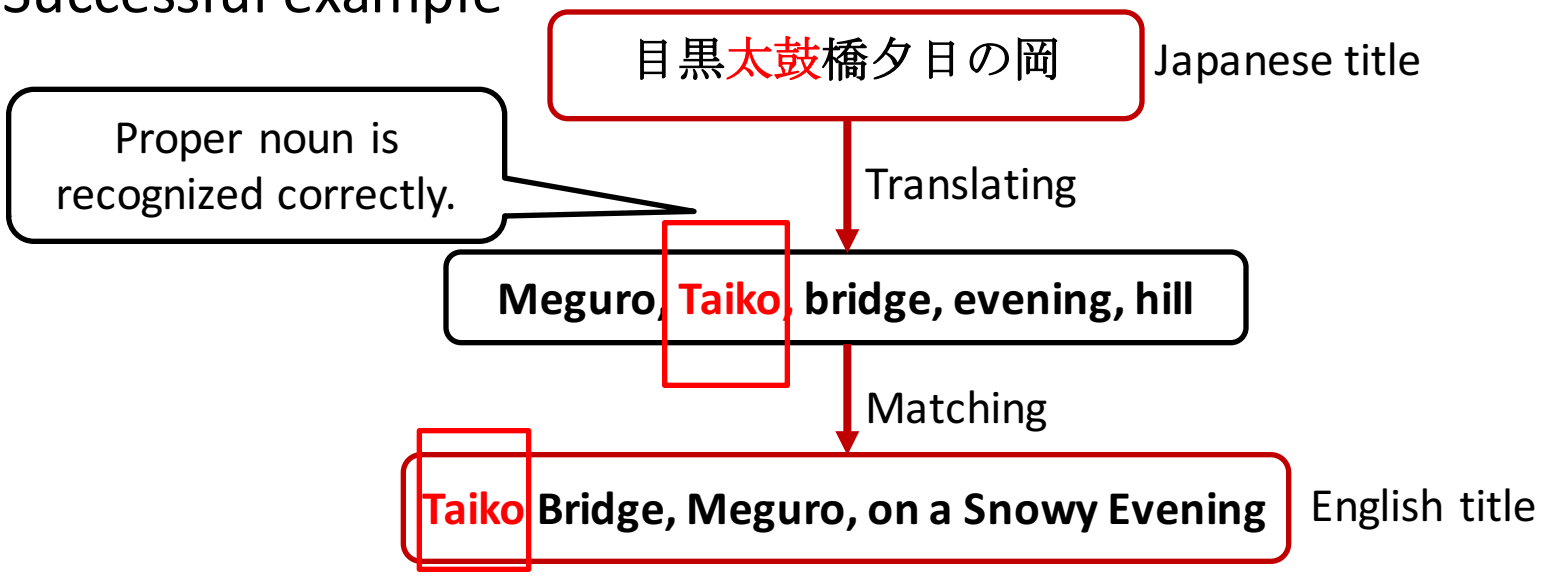
# Experimental results (2)

- Linking the same Ukiyo-e records that have **target titles**
  - Target titles contain at least one proper noun

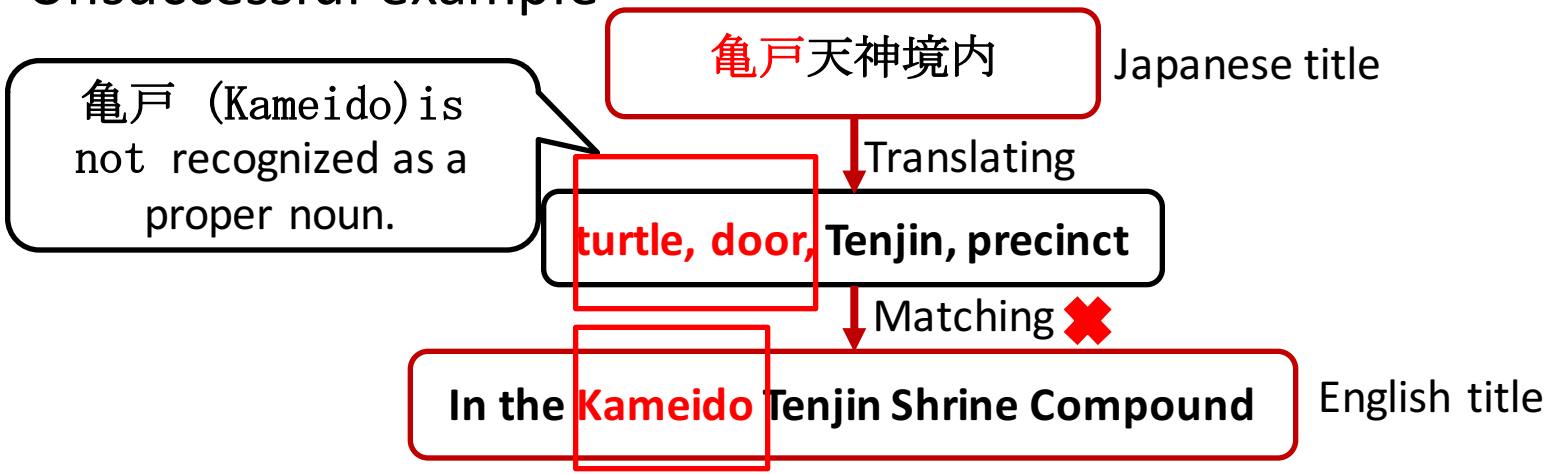


# Discussion

- Successful example




- Unsuccessful example



# Conclusion

- Recognizing and transliterating the proper nouns in cross-language record linkage effectively
- In the future, we plan to extend our method to classify the named entity type of acquired proper nouns

# Thank you!